Penerapan *Algoritma K-Means Clustering* Untuk Pengelompokkan Penyebaran Diare Di Kabupaten Langkat

Fina Nasari[1], Charles Jhony Manto Sianturi[2]
[1,2]Program Studi Sistem Informasi, Universitas Potensi Utama
E-mail : : finanasari@gmail.com[1], lapetgadong@yahoo.com[2]

Abstrak

Diare merupakan penyakit yang bertanggung jawab untuk sekitar seperempat dari 130.000 kematian tahunan diantara anak balita, terutama pada musim pancaroba seperti yang terjadi dihampir seluruh kawasan Indonesia tidak terlebih di kabupaten langkat sumatera utara. Untuk melihat kawasan penyebarannya perlu dibuat sebuah pengelompokan kawasan penyebaran diare, agar diperoleh daerah penyebaran diare dan pusat penyebarannya. Algoritma K-Means Clustering merupakan salah satu algoritma yang mengelompokkan data yang sama pada kelompok tertentu dan data yang berbeda pada kelompok yang lain. Hasil dari pengelompokkan daerah penyebaran diperoleh Kecamatan Batang Serangan, Brandan Barat dan Permata Jaya sebagai pusat penyebaran diare pada Cluster pertama dan Kecamatan Hinai dan Sei Bingai menjadi pusat cluster kedua.

Kata Kunci: Diare, Data Mining, K-Means Clustering

Abstract

Diarrhea is a disease that is responsible for about a quarter of the 130,000 annual deaths among children under five , especially in the transition season as happens in almost all Indonesian regions especially not in Langkat district of North Sumatra . To view the distribution area should be made a regional breakdown of the spread of diarrhea, in order to obtain the spread of diarrhea and regional distribution centers . K -Means Clustering Algorithm is one algorithm which classifies the same data at particular groups and different data in the other group . The results obtained from the grouping area deployment District of Batang Serangan , Brandan Barat and Permata Jaya as a center for the spread of diarrhea in the first cluster and the District Hinai and Sei Bingai became the center of the second cluster .

Keywords: Diarrhea , Data Mining , K -Means Clustering

1. PENDAHULUAN

Diare adalah frekuensi buang air besar lebih dari 4 kali pada bayi dan lebih dari tiga kali pada anak. Konsistensi feses encer dapat berwarna hijau atau dapat pula bercampur lender darah atau lender saja. Angka tingkat kematian yang dirilis UNICEF september 2012 menunjukkan bahwa secara global sekitar 2.000 anak di bawah usia lima tahun meninggal setiap hari akibat penyakit diare. Dari jumlah tersebut sebagian besar atau sekitar 1.800 anak per hari meninggal karena penyakit diare karena kurangnya air bersih, sanitasi dan kebersihan dasar[1].

Diperkirakan insidensi diare 0,5-2/episode/orang/tahun ada di negara maju sedangkan di negara berkembang lebih dari itu. Di USA dengan penduduk sekitar 200 juta diperkirakan 99 juta penderita diare setiap tahunnya [5].

Penyebaran penderita diare yang merata di hampir seluruh kawasan indonesia, salah satunya di kabupaten langkat. Kabupaten langkat merupakan salah satu kabupaten terbesar di provinsi sumatera utara yang letaknya berbatasan dengan provinsi aceh. Luasnya wilayah kabupaten langkat memungkinkan perlunya sebuah pengelompokan wilayah penyebaran diare, pengelompokan wilayah penyebaran diare akan

mengahasilkan titik-titik pusat penyebaran diarePenyebaran penderita diare yang merata di hampir seluruh kawasan indonesia, salah satunya di kabupaten langkat. Kabupaten langkat merupakan salah satu kabupaten terbesar di provinsi sumatera utara yang letaknya berbatasan dengan provinsi aceh. Luasnya wilayah kabupaten langkat memungkinkan perlunya sebuah pengelompokan wilayah penyebaran diare, pengelompokan wilayah penyebaran diare akan mengahasilkan titik-titik pusat penyebaran diare.

Datamining merupakan proses menemukan korelasi baru yang bermanfaat, pola dan *trend* dengan menambang sejumlah repositori data dalam jumlah besar, menggunakan teknologi pengenalan pola seperti statistik dan teknik matematika. *Datamining* semakin menyebar dan berkembang dengan pesat belakangan ini karena kemampuannya dalam menambang pola bermanfaat dan *trend* dari basis data yang sudah ada. Perusahaan-perusahaan telah menghabiskan dana milyaran untuk mengumpulkan data dalam jumlah *megabytes* atau *terabytes* tapi tidak mendapatkan keuntungan yang bernilai di dalamnya, padahal didalamnya terdapat informasi yang berharga namun tersembunyi pada repositori data [3].

Salah satu teknik pengelompokkan dalam data mining adalah metode *clustering*. Pengertian *clustering* keilmuan dalam data mining adalah pengelompokan sejumlah data atau objek ke dalam *cluster* (*group*) sehingga setiap dalam *cluster* tersebut akan berisi data yang semirip mungkin dan berbeda dengan objek dalam *cluster* yang lainnya[2].

Terdapat enam fungsi dalam data mining, yaitu:

1. Fungsi deskripsi (*description*)

Terkadang peneliti dan analis secara sederhana ingin mencoba mencari cara untuk menggambarkan pola dan kecenderungan yang terdapat dalam data. Sebagai contoh, petugas pengumpulan suara mungkin tidak dapat menemukan keterangan atau fakta bahwa siapa yang tidak cukup profesional akan sedikit didukung dalam pemilihan presiden. Deskripsi dari pola dan kecenderungan sering memberikan kemungkinan penjelasan untuk suatu pola atau kecenderungan.

2. Fungsi estimasi (*estimation*)

Estimasi hampir sama dengan klasifikasi, kecuali variabel target estmasi lebih ke arah numerik daripada ke arah kategori. Model dibangun menggunakan record lengkap yang menyediakan nilai dari variabel target sebagai nilai prediksi. Selanjutnya, pada penilaian berikutnya estimasi nilai dari variabel target dibuat berdasarkan nilai variabel prediksi. Sebagai contoh, akan dilakukan estimasi tekanan darah sistolik pada pasien rumah sakit berdasarkan umur pasien, jenis kelamin, indeks berat badan, dan level sodium darah. Hubungan antara tekanan darah sistolik dan nilai variabel prediksi dalam proses pembelajaran akan menghasilkan model esrimasi. Model estimasi yang dihasilkan dapat digunakan untuk kasus baru lainnya.

3. Fungsi prediksi (*prediction*)

Prediksi hampir sama dengan klasifikasi dan estimasi, kecuali dalam prediksi nilai dari hasil akan ada di masa mendatang. Contoh prediksi dalam bisnis dan penelitian adalah:

- Prediksi harga beras dalam tiga bulan mendatang.
- Prediksi presentasi kenaikan kecalakaan lalu lintas tahun depan jika batas bawah kecepatan dinaikkan.

4. Fungsi klasifikasi (*clasification*)

Di dalam klasifikasi terdapat target variabel kategori. Sebagai contoh penggolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu pendapatan tinggi,

pendapatan sedang, dan pendapatan rendah. Kemudian untuk menentukan pendapatan seorang pegawai, dipakai cara klasifikasi dalam data mining.

5. Fungsi pengelompokan (*clasification*)

Pengklusteran merupakan pengelompokkan record, pengamatan atau memperhatikan dan membentuk kelas objek-objek yang mempunyai kemiripan. Kluster adalah kumpulan record yang memiliki kemiripan satu dengan yang lainnya dan memiliki ketidakmiripan dengan record-record dalam kluster lain.

6. Fungsi asosiasi (*association*)

Tugas asosiasi dalam data mining adalah menemukan atribut yang muncul dalam satu waktu. Dalam dunia bisnis lebih umum disebut analisis keranjang belanja. Asosiasi mencari kombinasi jenis barang yang akan terjual untuk bulan depan [3].

Perbandingan metode pengelompokkan data, diperoleh kesimpulan bahwa metode pengelompokan atau klasifikasi data akan lebih optimal jika algoritma *K-Means* digabungkan dengan algoritma *hierarchical clustering. K-means* mempunyai kelemahan yang diakibatkan oleh penentuan pusat awal *cluster*. Hasil *cluster* yang terbentuk dari metode *K-means* ini sangatlah tergantung pada inisiasi nilai pusat awal *cluster* yang diberikan [2].

Perkembangan dari penerapan metode *k-means* diantaranya dalam pemilihan *distance space*, cara pengalokasian ulang data ke cluster dan *objective function* yang digunakan. *K-Means* juga telah dikembangkan untuk bisa memodel *dataset* yang mempunyai bentuk khusus dengan memanfaatkan kernel trik, Ada beberapa permasalahan yang perlu untuk diperhatikan dalam menggunakan metode *K-Means* termasuk model *clustering* yang berbeda-beda, pemilihan model yang paling tepat untuk *dataset* yang dianalisa, kegagalan untuk *converge*, pendeteksian *outliers*, bentuk masing-masing cluster dan permasalahan *overlapping*[4].

2. METODE PENELITIAN DAN HASIL PENELITIAN

2.1 Citra Pemilihan data (Data Selection)

Pada tahap pemilihan data, data yang dipakai adalah data jumlah penderita diare pada kecamatan-kecamatan yang ada di kabupaten langkat. Data penderita diare kabupaten langkat dapat dilihat pada Tabel. 1

Tabel 1. Data Penderita Diare Di Kabupaten Langkat

No.	Kecamatan	Luas (Km^2)	Jumlah Desa	Ibukota Kecamatan	Jumah Penderita Diare
1	Binjai	42,05	7	Kwala Begumit	24624
2	Sawit Seberang	209,10	7	Sawit Seberang	15277
3	Brandan Barat	89,80	7	Tangkahan Durian	10993
4	Kutambaru	236,84	8	Kutambaru	18131
5	Batang Serangan	899,38	8	Batang Serangan	11308
6	Babalan	76,41	8	Pelawi Utara	30298

No.	Kecamatan	Luas (Km^2)	Jumlah Desa	Ibukota Kecamatan	Jumah Penderita Diare
7	Pematang Jaya	209,00	8	Pematang Jaya	6943
8	Besitang	720,74	9	Pekan Besitang	9569
9	Sirapit	98,50	10	Serapit	20861
10	Gebang	178,49	11	Pekan Gebang	18565
11	Pangkalan Susu	151,35	11	Pangkalan Susu	5850
12	Stabat	108,85	12	Stabat Baru	17717
13	Padang Tualang	221,14	12	Tanjung Selamat	5668
14	Hinai	105,26	13	Tanjung Beringin	35947
15	Selesai	167,73	14	Pekan Selesai	17085
16	Wampu	194,21	14	Bingai	18550
17	Sei Lepan	280,68	14	Alur Dua	28135
18	Sei Bingai	333,17	16	Namu Ukur	20428
19	Kuala	206,23	16	Pekan Kuala	19183
20	Salapian	221,73	17	Minta Kasih	21094
21	Secanggang	231,19	17	Hinai Kiri	28515
22	Bahorok	1 101,83	19	Pekan Bahorok	20365
23	Tanjung Pura	179.61	19	Pekan Tanjung	17396

2.2 Transformasi Data (Pra Processing)

Proses Transformasi Data atau Pra Processing adalah proses perubahan data menjadi data yang dapat diolah menggunakan algoritma yang akan dipakai apakah dalam bentuk numerik, klasifikasi dan lain-lain. Kriteria transformasi data diare dapat dilihat pada Tabel 2 untuk tranformasi jumlah penderita diare dan Tabel 3 untuk kriteria transformasi data kecamatan. Hasil transformasi data dapat dilihat pada Tabel 4.

Tabel 2 Kriteria Transformasi Data Jumlah Penderita Diare

Kriteria Transformasi Jumlah Penderita Diare	
Jumlah Penderita	**Transformasi**
<=1000 Jiwa	1
1001 s/d 5000 Jiwa	2
50001 s/d 10000 Jiwa	3
10001 s/d 15000 Jiwa	4

Kriteria Transformasi Jumlah Penderita Diare	
Jumlah Penderita	**Transformasi**
15001 s/d 20000 Jiwa	5
>20000 Jiwa	6

Tabel 3 Kriteria Transformasi Data Kecamatan

Kriteria Transformasi Kecamatan	
Jumlah Desa	**Transformasi**
1 s/d 5 Desa	1
6 s/d 10	2
11 s/d 15	3
16 s/d 20	4
21 s.d 25	5
>25	6

Tabel 4 Hasil Transformasi Data Penderita Diare Kabupaten Langkat

No.	Kecamatan	Jumlah Penderita	Transformasi Kecamatan	Transformasi Jumlah Penderita
1	Babalan	30298	2	6
2	Bahorok	20365	4	6
3	Batang Serangan	11308	2	4
4	Besitang	9569	2	3
5	Binjai	24624	2	6
6	Brandan Barat	10993	2	4
7	Gebang	18565	3	5
8	Hinai	35947	3	6
9	Kuala	19183	4	5
10	Kutambaru	18131	2	5
11	Padang Tualang	5668	3	3
12	Pangkalan Susu	5850	3	3

No.	Kecamatan	Jumlah Penderita	Transformasi Kecamatan	Transformasi Jumlah Penderita
13	Pematang Jaya	6943	2	3
14	Salapian	21094	4	6
15	Sawit Seberang	15277	2	5
16	Secanggang	28515	4	6
17	Sei Bingai	20428	3	6
18	Sei Lepan	28135	4	6
19	Selesai	17085	3	5
20	Sirapit	20861	2	6
21	Stabat	17717	3	5
22	Tanjung Pura	17396	4	5
23	Wampu	18550	3	5

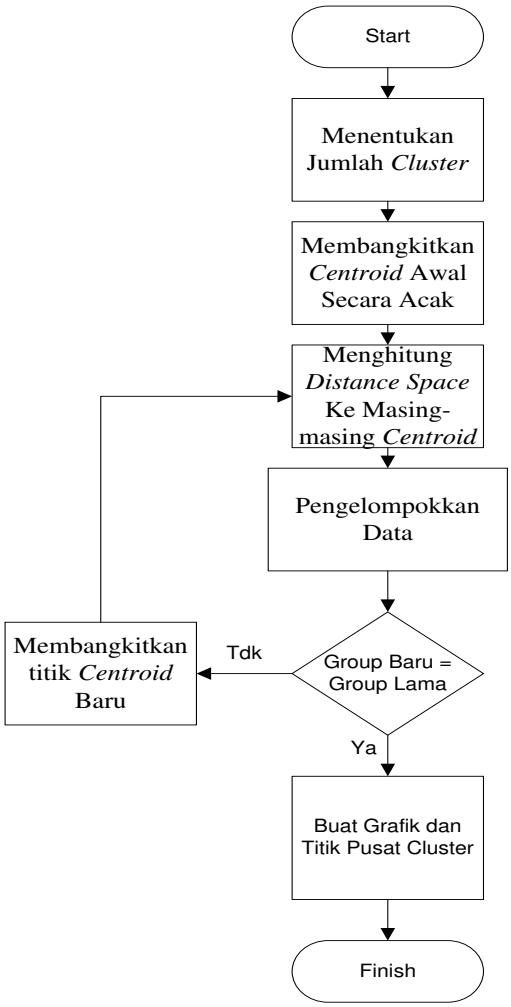
2.3 Penerapan Algoritma K-Means Clustering Untuk Penyebaran Diare

Tahapan Algoritma K-Means Clustering dapat dilihat pada gambar 1.

1. Menentukan jumlah cluster, jumlah cluster merupakan jumlah kelompok data yang akan dibuat atau dihasilkan. Dalam penelitian ini jumlah cluster yang akan dibuat adalah 2 cluster.
2. Membangkitkan centroid awal. Centroid awal diperoleh secara acak, dan jumlah centroid sebanyak cluster yang akan dibuat. Centroid awal merupakan titik pusat cluster pertama atau awal pusat cluster. Controid awal dari penelitian ini adalah:
   Centroid-1 (4.00, 6.00)
   Centroid-2 (2.00, 3.00)
3. Menghitung distance space data ke masing-masing centroid.
   Formula Menghitung jarak ke masing-masing cluster

$$d_{euclidean}\ (x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad \text{.....[3]}$$

Dimana:
x dan y : representasi nilai atribut dari dua record .



Gambar 1 Tahapan Algoritma K-Means Clustering

Menghitung jarak data ke masing-masing centroid, hasil perhitungan jarak ke masing-masing centroid dapat dilihat pada Tabel 5

Tabel 5 Hasil Perhitungan jarak ke masing-masing *centroid*

No.	Transformasi Kecamatan	Transformasi Jumlah Penderita	Jarak Ke Centroid 1	Jarak Ke Centroid 2
1	2	6	2.000	3.000
2	4	6	0.000	2.236
3	2	4	0.000	1.000
4	2	3	2.236	0.000
5	2	6	2.000	3.000
6	2	4	0.000	1.000
7	3	5	0.000	1.732
8	3	6	1.000	2.828
9	4	5	1.000	0.000

No.	Transformasi Kecamatan	Transformasi Jumlah Penderita	Jarak Ke Centroid 1	Jarak Ke Centroid 2
10	2	5	1.732	2.000
11	3	3	2.828	1.000
12	3	3	2.828	1.000
13	2	3	2.236	0.000
14	4	6	0.000	2.236
15	2	5	1.732	2.000
16	4	6	0.000	2.236
17	3	6	1.000	2.828
18	4	6	0.000	2.236
19	3	5	0.000	1.732
20	2	6	2.000	3.000
21	3	5	0.000	1.732
22	4	5	1.000	0.000
23	3	5	0.000	1.732

4. Pengelompokkan data *cluster.*

Hasil perhitungan *distance space* atau jarak ke masing-masing *centroid*, langkah selanjutnya adalah mengelompokkan data. Data tersebut akan dikelompokkan pada centoid pertama, kedua atau ketiga. Hasil Pengelompokkan iterasi pertama dapat dilihat pada Tabel 6

Tabel 6 Hasil Pengelompokkan Data Pada Iterasi pertama

No.	Transformasi Kecamatan	Transformasi Jumlah Penderita	Jarak Ke Centroid 1	Jarak Ke Centroid 2	Group Awal	Group Baru
1	2	6	2.00	3.00	0	1
2	4	6	0.00	2.24	0	1
3	2	4	0.00	1.00	0	1
4	2	3	2.24	0.00	0	2
5	2	6	2.00	3.00	0	1
6	2	4	0.00	1.00	0	1
7	3	5	0.00	1.73	0	1
8	3	6	1.00	2.83	0	1
9	4	5	1.00	0.00	0	2
10	2	5	1.73	2.00	0	1
11	3	3	2.83	1.00	0	2
12	3	3	2.83	1.00	0	2
13	2	3	2.24	0.00	0	2
14	4	6	0.00	2.24	0	1
15	2	5	1.73	2.00	0	1

No.	Transformasi Kecamatan	Transformasi Jumlah Penderita	Jarak Ke Centroid 1	Jarak Ke Centroid 2	Group Awal	Group Baru
16	4	6	0.00	2.24	0	1
17	3	6	1.00	2.83	0	1
18	4	6	0.00	2.24	0	1
19	3	5	0.00	1.73	0	1
20	2	6	2.00	3.00	0	1
21	3	5	0.00	1.73	0	1
22	4	5	1.00	0.00	0	2
23	3	5	0.00	1.73	0	1

5. Kembali ke Step 3, apabila masih ada data yang berpindah cluster atau apabila ada perubahan nilai *centroid*, ada yang di atas nilai *threshold* yang ditentukan atau apabila perubahan nilai pada *objective function* yang digunakan di atas nilai *threshold* yang ditentukan [2].

Sedangkan formula membangkitkan *cendroid* baru

$$C = \frac{\sum m}{n} \quad \text{.....[3]}$$

Dimana:

C : *centroid* data
m : anggota data yang termasuk kedalam *centroid* tertentu
n : jumlah data yang menjadi anggota *centroid* tertentu

Hasil pengelompokkan data pada iterasi pertama berbeda dengan kelompok awal, oleh karena itu lanjut pada tahap berikutnya yaitu menentukan kembali nilai *centroid* atau titik pusat data sementara.

Membangkitkan *centroid* baru berdasarkan group baru yang dihasilkan:

$$C = \frac{\sum m}{n} \quad \text{.....[3]}$$

Centroid 1: (48/17 ; 92/17) → (2.82 ; 5.41)
Centroid 2: (18/6 ; 22/6) → (3.00 ; 3.67)

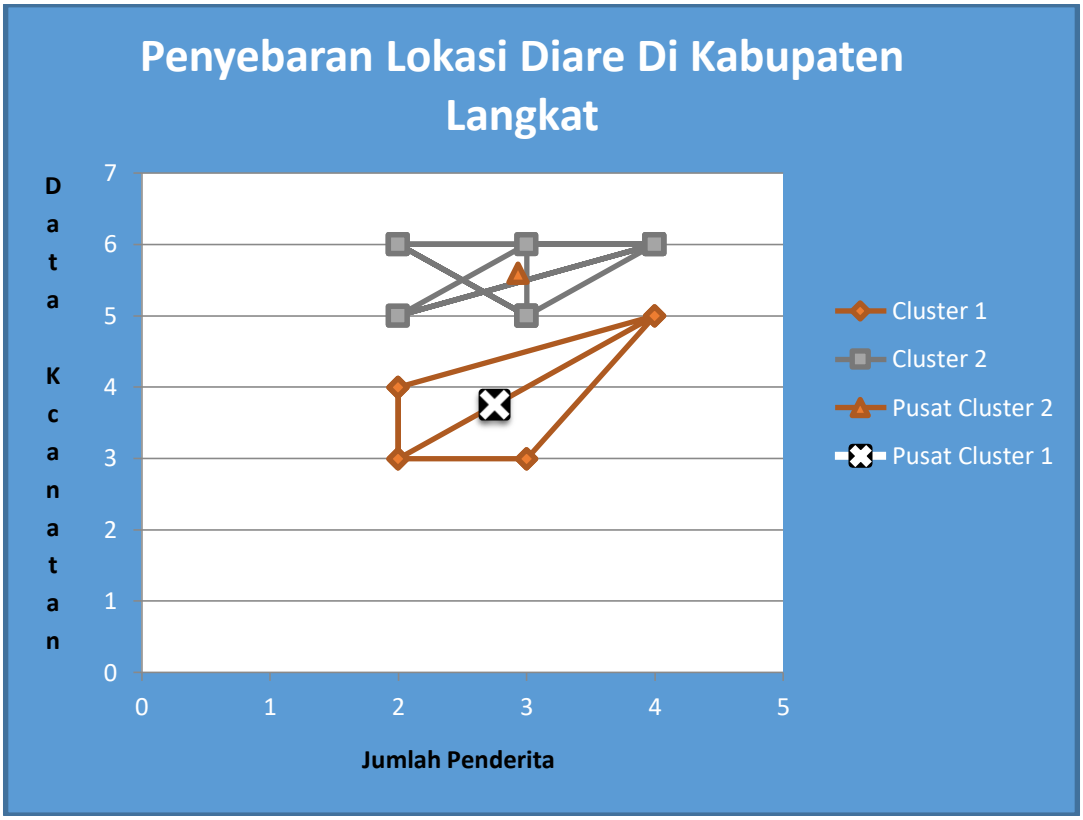
6. Menghitung *distance space* ke *centorid* yang baru untuk iterasi berikutnya.

Dengan menggunakan nilai *cetroid* yang baru, jarak ke masing-masing *centroid* dihitung sampai group baru sama dengan group yang lama atau group sebelumnya. Hasil Penerapan algoritma *K- Means Clustering* dapat dilihat pada Tabel 7

Tabel 7 Hasil Penerapan Algoritma *K-Means* Clustering

No.	Kecamatan	Jumlah Penderita	Transformasi Kecamatan	Transformasi Jumlah Penderita	Jarak Ke Centroid 1	Jarak Ke Centroid 2	Group Awal	Group Baru
1	**Batang Serangan**	**11308**	**2**	**4**	**0.71**	**1.30**	**1**	**1**
2	Besitang	9569	2	3	0.00	2.43	1	1
3	**Brandan Barat**	**10993**	**2**	**4**	**0.71**	**1.30**	**1**	**1**
4	Kuala	19183	4	5	0.00	0.89	1	1
5	Padang Tualang	5668	3	3	0.71	2.60	1	1
6	Pangkalan Susu	5850	3	3	0.71	2.60	1	1
7	**Pematang Jaya**	**6943**	**2**	**3**	**0.00**	**2.43**	**1**	**1**
8	Tanjung Pura	17396	4	5	0.00	0.89	1	1
9	Babalan	30298	2	6	2.12	0.84	2	2
10	Bahorok	20365	4	6	1.87	0.99	2	2
11	Binjai	24624	2	6	2.12	0.84	2	2
12	Gebang	18565	3	5	1.22	0.60	2	2
13	**Hinai**	**35947**	**3**	**6**	**2.24**	**0.39**	**2**	**2**
14	Kutambaru	18131	2	5	1.00	0.71	2	2
15	Salapian	21094	4	6	1.87	0.99	2	2
16	Sawit Seberang	15277	2	5	1.00	0.71	2	2
17	Secanggang	28515	4	6	1.87	0.99	2	2
18	**Sei Bingai**	**20428**	**3**	**6**	**2.24**	**0.39**	**2**	**2**
19	Sei Lepan	28135	4	6	1.87	0.99	2	2
20	Selesai	17085	3	5	1.22	0.60	2	2
21	**Sirapit**	**20861**	**2**	**6**	**2.12**	**0.84**	**2**	**2**
22	Stabat	17717	3	5	1.22	0.60	2	2
23	Wampu	18550	3	5	1.22	0.60	2	2

Berdasarkan Tabel 7 diperoleh titik pusat ***cluster* 1 : (2.75 ; 3.75)** atau pusat *cluster* berada pada kecamatan Batang serangan, Brandan Barat dan pematang jaya, sedangkan pusat ***cluster 2 :* (2.93 ; 3.60)** berada pada kecamatan Hinai, Sei Bingai dan Sirapit. Group terakhir yang dihasilkan selanjutnya digambarkan dalam sebuah grafik *Cluster* data dengan nilai *centroid* terkahir menjadi titik pusat *cluster.* Grafik hasil penerapan algoritma *K-Means Clustering* dapat dilihat pada Gambar 2.



Gambar 2 Hasil penerapan algoritma *K-Means Clustering*

3. KESIMPULAN DAN SARAN

Kesimpulan dari penelitian ini adalah:

1. Pusat *cluster* yang diperoleh yaitu untuk *cluster* pertama berada pada kecamatan Batang serangan, Brandan Barat dan Pematang Jaya dan pusat *cluster* kedua berada pada kecamatan Hinai, Sei Bingai dan Sirapit.
2. Pusat *cluster* pertama merupakan daerah penyebaran diare untuk jumlah penderita tingkat menengah atau bukan merupakan pusat penyebaran diare.
3. Pusat *cluster* kedua merupakan daerah-daerah pusat penyebaran diare, untuk itu pada daerah-daerah pusat *cluster* kedua harus menjadi daerah perhatian pemerintah untuk penanganan diare.

Saran Dalam Penelitian Ini adalah:

1. Sebaiknya untuk penelitian selanjutnya menggunakan data perkelurahan agar daerah-daerah penyebaran diperoleh lebih detail.
2. Sebaiknya dalam penentuan *cluster* pertama dibantu dengan algoritma tertentu agar hasil *cluster* yang diperoleh lebih optimal.

DAFTAR PUSTAKA

[1]. http://www.unicef.org/indonesia/id/media_19772.html

[2]. Alfina, Tahta et. al, 2012, Analisa perbandingan metode hierarchical clustering, k means dan gabungan keduanya dalam cluster data (studi kasus : problem kerja praktek jurusan teknik industri its). Surabaya: jurnal teknik its vol. 1, (sept, 2012) issn: 2301-9271

[3]. Larose, daniel . 2005. Discovery knowledge in data, a jhon wiley & sons, inc publication. Canada

[4]. Agusta yudi, 2007, k-means-penerapan, permasalahan dan metode terkait. Denpasar, bali: 2007 jurnal sistem dan informatika vol. 3 (pebruari 2007), 47-60

[5]. Andyanastri festy, 2012, etiologi dan gambaran klinis diare akut di rsup dr kariadi semarang. Semarang, jurnal ilmiah kti